修士論文の和文要旨

大学院情報システム学研究科	博士前期課程	情報システム設計学 専攻
氏名	矢部 走	学籍番号 0650035
論文題目	文書構造化を用いたWebページからのRSS自動生成システム	

要旨

多くのWebページから必要な情報を効率的に取得するための手段としてRSSが注目を集めている．RSSはWebサイトの各ページのタイトル，URL，概要，更新時刻などを構造化して記述するXMLベースのフォーマットで，Webページの更新情報を配信するのに適しており，多くのサイトで提供され始めている．しかし，現状ではRSSを配信していないWebページも多くRSSの利用の弊害となっている．その問題を解決するため，WebページからRSSを自動生成するサービスもいくつか存在する．それらのサービスで用いられているアルゴリズムは日付抽出とリンク抽出の2つに大別できるが，どちらも対応するページの種類が限られており，RSSを生成することが出来ない場合も多い．特に，テキスト主体のWebページおけるRSSの生成は難しいのが現状である．

そこで本論文では，テキスト主体のWebページからもRSSを生成すべく，既存の手法である文書構造化，そして日付表現を用いたタイトル抽出を用い，さらにそれに独自の要素としてHTML文書をグループ構造化し，そのグループ構造とDPマッチングを用いてタイトルの抽出を行う方法を組み合わせたRSS生成手法を提案し，従来RSS情報の抽出が困難であったWebページからのRSS自動生成を試みる．

実験では18種類，153のページを対象に既存手法と提案手法それぞれによるRSS生成を行い，精度約65％の既存手法に対し，提案手法は約85％の精度でRSSを生成できることを示した．さらに提案手法は，既存の手法ではRSS生成が難しかったテキスト系のWebページからのRSS生成に対応し，他の種類のWebページよりも精度が高いという結果を残した．